W186–W192 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 4 June 2012*
doi:10.1093/nar/gks528

VIGOR extended to annotate genomes for additional 12 different viruses

Shiliang Wang[1,*], Jaideep P. Sundaram[2] and Timothy B. Stockwell[1]

[1]J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850 and [2]Genomics Department, BioReliance Corporation, 14920 Broschart Rd, Rockville, MD 20850, USA

Received March 8, 2012; Revised April 27, 2012; Accepted May 11, 2012

ABSTRACT

A gene prediction program, VIGOR (Viral Genome ORF Reader), was developed at J. Craig Venter Institute in 2010 and has been successfully performing gene calling in coronavirus, influenza, rhinovirus and rotavirus for projects at the Genome Sequencing Center for Infectious Diseases. VIGOR uses sequence similarity search against custom protein databases to identify protein coding regions, start and stop codons and other gene features. Ribonucleicacid editing and other features are accurately identified based on sequence similarity and signature residues. VIGOR produces four output files: a gene prediction file, a complementary DNA file, an alignment file, and a gene feature table file. The gene feature table can be used to create GenBank submission. VIGOR takes a single input: viral genomic sequences in FASTA format. VIGOR has been extended to predict genes for 12 viruses: measles virus, mumps virus, rubella virus, respiratory syncytial virus, alphavirus and Venezuelan equine encephalitis virus, norovirus, metapneumovirus, yellow fever virus, Japanese encephalitis virus, parainfluenza virus and Sendai virus. VIGOR accurately detects the complex gene features like ribonucleicacid editing, stop codon leakage and ribosomal shunting. Precisely identifying the mat_peptide cleavage for some viruses is a built-in feature of VIGOR. The gene predictions for these viruses have been evaluated by testing from 27 to 240 genomes from GenBank.

INTRODUCTION

The advance of sequencing technologies, especially next-generation sequencing technologies, makes genome sequencing more efficient and more cost-effective. Tremendous quantities of sequence data, particularly the sequences of clinical specimens, are being generated daily by sequencing centres and individual laboratories. The prediction and annotation of genes is the first and most critical step in understanding the structure and function of genome components of viruses, bacteria and eukaryotes. However, the development of tools that interpret sequence data still lags in pace of sequence data production.

Investigators from Microbial Sequencing Center (MSC, 2003) and Genome Sequencing Center for Infectious Diseases (GSCID, 2009) at J. Craig Venter Institute, sponsored by the National Institute of Allergy and Infectious Diseases, are sequencing various pathogenic genomes, including different viral genomes. Thousands of influenza genomes and hundreds of coronavirus genomes, rotavirus genomes and other viral genomes were sequenced by MSC. Viral genomes are relatively small (ranging from a few kb to 100 kb) compared with bacterial and eukaryotic genomes. However, the genome structure and gene features can be as complex as eukaryotic genomes. In addition to the normal splicing to generate mature messenger ribonucleicacid (mRNA) like the splicing in eukaryotic genome, a variety of complex gene features, such as alternative splicing, ribosomal slippage, ribonucleicacid (RNA) editing, stop codon leakage (stop codon read-through), ribosomal shunting (alternative translation initiation) and gene overlapping, can also be found in many viral genomes. A large number of viruses, such as rhinovirus and yellow fever virus, have only one open reading frame encoded in their small genomes. The single polyprotein of these viruses needs to be processed into mature peptides that will be incorporated into virus structure and be functional during virus replication. These complex gene features must be accurately defined in order to correctly understand these genomes.

Universal gene prediction programs, such as FgenesV (www.softberry.com) and Zcurve_V (1), are available for public use. However, accurate detection of the complex gene features is a major hurdle for these universal gene prediction tools. Therefore, species-specific gene prediction

*To whom correspondence should be addressed. Tel: +1 301 795 7842; Fax: +1 301 294 3142; Email: shiliang@jcvi.org

programs have been developed to conduct the gene calling for specific viral genomes. For example, FLAN was developed at NCBI to annotate influenza viral genomes (2). In order to efficiently validate genome sequence data and accurately annotate the viral genomes sequenced by MSC, a homology-based viral gene prediction program called VIGOR (Viral Genome ORF Reader, http://www.jcvi.org/vigor) was developed at J. Craig Venter Institute (JCVI) for influenza virus, coronavirus, SARS coronavirus, rhinovirus and rotavirus (3).

In addition to the viruses listed earlier, measles virus, mumps virus, rubella virus, norovirus, metapneumovirus (MPV), respiratory syncytial virus (RSV), human parainfluenza virus 1 (hPIV-1) and 3 (hPIV-3), Venezuelan equine encephalitis virus (VEEV), yellow fever virus (YFV) and Japanese encephalitis virus (JEV) are also being sequenced by GSCID at JCVI. In order to validate and accurately annotate these genomes, viral species-specific protein databases have been created and VIGOR's capabilities were enhanced to detect the new species-specific gene features, such as RNA editing and stop codon read-through. The accuracy of gene predictions of this enhanced version has been evaluated against the same viral annotations available at NCBI. VIGOR was also extended to predict genes correctly for related viruses, such as alphavirus and Sendai virus.

METHODS

The processes for collecting sequences to create protein databases, detecting protein coding regions, identifying start and stop codons, predicting mature peptide cleavage sites, as well as a definition of correct gene prediction and an overview of the program implementation are outlined in an article by Wang *et al.* (3). What follows is a discussion of the methods developed to address the gene features of the new viruses.

RNA editing

In many RNA viruses, such as measles virus, mumps virus and parainfluenza virus, the P gene is a polycistronic gene from which multiple proteins can be produced. For measles virus, hPIV-3, bovine parainfluenza virus 3 (bPIV-3) and Sendai virus, the faithful transcript of the P gene will guide synthesis of the P protein; the edited mRNA, in which one or more non-templated G residues are inserted, will be translated into the V or D protein. To identify the RNA editing site, reference V or D protein sequence is aligned to the protein sequence deduced from genomic sequence to locate the frame-shifting region. In measles virus, Sendai virus, bPIV-3, hPIV-3, the sequence string AAAAAGGG immediately upstream the editing site is conserved and is screened to identify the RNA editing site (4,5). A single G residue is added into the complementary DNA sequence of measles virus at the editing site and the reading frame shifts from +1 to +2; in hPIV-3, bPIV-3 and Sendai virus, two G residues are inserted at the editing site. V protein sequence will be deduced from this edited mRNA sequence. In mumps virus, hPIV-2 and hPIV-4, the faithful transcript encodes V protein, and the edited mRNA, in which two non-templated G residues are added, encodes the P protein. The conserved nucleotide sequence string upstream the editing site is AAGAGGG (6–8).

Stop codon leakage

The genomes of VEEV and alphavirus encode only two genes. However, two different proteins can be produced from the first large gene by a mechanism known as stop codon leakage (also known as stop codon read-through) (9). For the majority of mRNA, translation will terminate when the ribosome reaches the first stop codon (UGA). An Arg residue, however, will be incorporated into the protein product at the stop codon position (UGA) for a portion of mRNA (5–50%), and the translation will be aborted at the next stop codon downstream. The stop codon read-through site region is selected by aligning the deduced amino acid sequence to the reference protein in database, and confirmed by the identity (>60%) of the peptide sequences flanking the potential stop leakage site (100 residues on each side). The leakage site is finalized by the codon (UGA) and the conserved motif (Arg-Leu) next to the read-through site (9).

Ribosomal shunting

Within the P gene region of parainfluenza viruses and Sendai virus, a maximum of four proteins (C', C, Y1 and Y2) are encoded in addition to the P, V and D proteins by a mechanism known as ribosomal shunting (also known as alternative translation initiation) (10). The start codon of C protein gene is downstream the start codon of P protein gene, but the reading frame of C protein gene is different. VIGOR predicts the C protein coding sequence first. The next two in-frame ATG codons in the C protein coding sequence will be the start codons of the Y1 and Y2 protein genes (11). The start codon (ACG or GCG) of C' protein will be scanned upstream of the P protein start codon (7 to 15 nt) (12), but in the same reading frame as the C protein gene.

Implementation

The VIGOR interface was implemented in Perl CGI, PHP, HTML, CSS and Javascript, and the executables for gene prediction were programmed by Perl as described (3). The VIGOR interface has been adjusted to accommodate some new features. Users can submit the FASTA or multi-FASTA format genomic sequences by pasting them into the input sequence area or uploading a local sequence file. The user e-mail address is optional and needs only be supplied if the user wishes to receive a notification e-mail from the system. A URL for the result page will be displayed immediately on submitting data for prediction. The time needed to complete the gene predictions and format the results depends on the size of data submitted by the user. Users can periodically refresh the result page or wait for the notification e-mail, if it is provided. Either way, users can bookmark the result page for future reference.

In addition to the gene prediction file, a complementary DNA file and an alignment file as described (3), a gene

feature table (the TBL link in the result page) is also generated by this enhanced VIGOR. The gene feature table can be used to create GenBank submission files by Sequin (http://www.ncbi.nlm.nih.gov/Sequin/index) or tbl2asn (http://www.ncbi.nlm.nih.gov/GenBank/tbl2asn2.html). A downloadable tar file, including all the four output files, is also generated. Users can download it from the result page. A detail instruction can be found from the instruction page of VIGOR website.

EVALUATION AND DISCUSSION

Since no golden standard dataset is available for the viruses that VIGOR is designed to predict genes for and no species-specific gene prediction programs are available for the designated viruses, the performance of VIGOR gene prediction was evaluated by comparing VIGOR prediction with GenBank annotation of the same viral sequence. For each individual virus types, 27 to 240 complete genomes, depending on the availabilities in NCBI, were run through VIGOR for comparison. The comparison data is presented in Table 1. The full descriptions of correct prediction, partial agreement, missing gene and new gene were depicted previously (3). In brief, a prediction is considered correct if both start and stop codons of the gene predicted by VIGOR are the same as these in the GenBank annotation, and if the gene-specific features (like the RNA editing site and stop read-through site) exist, then they must agree with the GenBank record. The following two types of predictions will be counted as partial agreements: stop codon and the reading frame are the same as these in GenBank record, but start codon is different; for some genes, an internal stop codon is detected and the truncated protein sequence is shorter than 95% of its reference sequence length (protein length is >150 aa), VIGOR defines these genes as nonfunctional genes and marks the predictions with 'possible sequence mutation,' but these genes are annotated as functional genes in GenBank. Missing gene means that no gene is detected by VIGOR in a region where one or more genes are annotated in GenBank. If a gene is predicted in a genomic sequence and no gene was documented in the same region with same reading frame in GenBank, this prediction will be inspected manually. If it is highly homologous to a related viral protein (E < 1e-10), the prediction will be counted as a new gene.

YFV and JEV

YFV and JEV are single-stranded, positive sense RNA viruses and belong to viral family Flaviviridae. The genome sizes of these viruses are ~11 kb and only a single long polyprotein is encoded in these genomes. This long polyprotein will be co- and post-translationally processed into 10 to 11 mature peptides by host signal peptidases and a virus-encoded serine protease. The three mature peptides (C, PrM/M and E) from the N-terminus of the polyprotein are structural proteins, and the other seven or eight mature peptides from the C-terminus are non-structural proteins. One hundred seventy complete genomes of both YFV and JEV were tested by VIGOR, a total of 170 polyprotein genes were predicted by VIGOR because only a single gene is encoded in each genome (13,14). The start and stop codons of these genes were in complete agreement with GenBank annotations (Table 1). The polyprotein of JEV is cleaved into 10 mature peptides (13). Eleven mature peptides are encoded in YFV genome (14). VIGOR was able to identify the mature peptide cleavage sites accurately and predicted 10 and 11 mature peptides in the genomes of these two viruses. Only a small number of genomes of these two viruses have their mature peptides annotated in GenBank. We compared the available limited data with VIGOR predictions. VIGOR mature peptide predictions agreed with GenBank records (data not shown). The conserved signature sequences for cleavage sites were also observed in VIGOR predictions.

Norovirus

Norovirus is responsible for at least 23 million cases of foodborne illness every year in the USA (15). The spread of norovirus can result in gastroenteritis epidemics. Noroviruses constitute a genus in the calicivirus family. The viruses enclose a non-segmented, positive single-stranded RNA genome of ~7.7 kb with three genes encoded (16). The second and third genes encode viral structure proteins named VP1 and VP2. The first large ORF encodes a non-structural protein that will be digested into six mature peptides by the viral 3C-like

Table 1. Comparison data between VIGOR predictions and GenBank annotations of the designated viruses

Virus name	No. of genomes	No. of GenBank Annotations	No. of VIGOR predictions	No. of same predictions	Partial agreements	Missing gene	New gene	Special features	Mat_peptide prediction
MPV	27	243	243	235	8	0	0	Gene overlapping	N/A[a]
Measles and mumps viruses	101	759	774	754	5	0	15	RNA editing	N/A
Norovirus	98	294	294	294	0	0	0	No	Yes
Parainfluenza virus and Sendai virus	36	249	311	249	0	0	62	RNA editing, Ribosomal shunting	N/A
RSV	50	535	568	522	13	0	33	Gene overlapping	N/A
Rubella virus	46	92	92	91	1	0	0	No	Yes
VEEV and alphavirus	240	478	478	478	0	0	0	Stop codon leakage,	Yes
YFV and JEV	170	170	170	170	0	0	0	No	Yes

[a]Not applicable.

protease ($3CL^{pro}$). The mature peptides p48, NTPase, p22, VPg, $3CL^{pro}$ and RdRp are ordered from N- to C-terminus (16).

VIGOR predicted 294 genes in 98 norovirus genomes. All of these 294 predictions agreed with GenBank annotations (Table 1). Only a small number of norovirus genomes have their mature peptides annotated in GenBank. The mature peptides predicted by VIGOR completely agreed with GenBank annotations when available. For those without mature peptide annotations in GenBank, the conserved cleavage sites (p48-(QG)-NTPase-(QG)-p22-(EG)-VPg-(EA)-$3CL^{pro}$-(EA)-RdRp) were inspected, and the alignments between VIGOR prediction and homologous peptides were also examined. All the mature peptides predicted by VIGOR were correct (data not shown).

Metapneumovirus

MPV is a genus in the paramyxovirus family. Viruses in this genus are also enveloped, non-segmented, negative-strand RNA viruses. The genome of MPV is ~13 kb and encodes eight genes (N-P-M-F-M2-SH-G-L). Two ORFs (M2-1 and M2-2) are encoded in the M2 gene region with overlapping and different reading frames (17,18).

Twenty-seven complete MPV genomes were used to evaluate VIGOR. A total of 243 genes were detected by VIGOR, the same number of genes were annotated in GenBank. VIGOR predictions completely agreed with GenBank annotations for 235 genes. VIGOR detected internal stop codons in the coding regions of four genes and the truncated proteins were shorter than 95% of the reference protein length. VIGOR, therefore, categorized these four genes as pseudogenes. These genes were annotated as functional genes in GenBank. The start codons are the same as these predicted by VIGOR. The stop codons in GenBank of these genes are same as the internal stop codons detected by VIGOR. The other four cases, in which VIGOR gene predictions were not same as GenBank annotations, are the M2 ORF1 gene (M2-1). Two versions of M2-1 protein exist in GenBank protein database. The longer protein has five additional amino acids at the N-terminus. VIGOR always selects the longer protein to be the reference sequence. The upstream ATG was therefore selected as the start codon in these four genes. In the corresponding GenBank annotations, the downstream ATG were annotated as the start codons for these four genes.

Measles and mumps viruses

Measles virus has a non-segmented, negative-stranded RNA genome, which is ~15.8 kb and encodes six genes (N, P, M, F, H and L). In addition to the P protein, the P gene also encodes two other proteins (V and C) through different mechanisms. The C protein translation initiates from an alternative start codon with a different reading frame; the V protein is identical to the P protein at the N-terminal domain, but the C-terminal domain is different because of RNA editing, which results in one G residue insertion at the RNA editing site. In the mumps virus genome, one extra small hydrophobic protein is encoded between the fusion gene and the HN gene. The P gene encodes two proteins (P and V). Two non-templated G residues are inserted into the mature transcript that is going to be translated into the P protein, and the mRNA without editing encodes the V protein (6).

A total of 101 complete genomes of measles and mump viruses were downloaded from NCBI. VIGOR predicted 774 genes, whereas GenBank record indicates 759 genes in these genomes. VIGOR detected 15 new genes validated by comparing them with other viral proteins (data not shown here). Most of them are the C protein gene, which overlaps with P protein gene but in a different reading frame. Five VIGOR predictions have different start codons with the same reading frame as GenBank annotations. All these five predictions are the fusion protein gene. Like the M2-1 protein of MPV, two versions of fusion protein exist in the GenBank protein database. VIGOR selected the longer one as the reference sequence to define the start codon. In GenBank annotations, the downstream ATG was selected as the start codon.

VEEV and alphavirus

VEEV and alphavirus belong to the togavirus family, which are single-stranded, positive sense RNA viruses. The genomes of VEEV and alphavirus are ~11.5 kb, and the 5' two-thirds of RNA encodes two non-structural proteins, which are cleaved into three or four peptides required for RNA genome replication. The 3' one-third of the genomes encodes a structural protein that is cleaved into three mature peptides (C, E2 and E1) by proteases (9). In the first large gene of VEEV and some alphaviruses, a stop codon (UGA) is present downstream of the coding sequence of mature peptide nsP3. The translation stops at this UGA codon on most mRNA. However, for a small percentage of mRNA, translational read-through occurs at this stop codon, an arginine or another residue is incorporated at this position and the translation is extended to the next stop codon (9,19,20). A larger non-structural protein is synthesized because of the translational read-through and one additional mature peptide (nsP4) is translated. The mature peptide cleavage sites are identified based on the alignments between the polyprotein and the homologs in the mature peptide database. Once the junction region is located between two consecutive mature peptides, the conserved signature sequences next to the cleavage sites are used to identify the exact position of the cleavage. The conserved signature sequences were documented in literature (9).

A total of 478 genes in 240 VEEV and alphavirus genomes were detected by VIGOR, and all of them were the same as the annotations in GenBank (table 1). The stop codon read-through was also detected accurately by VIGOR and a longer protein was generated. Only a small number of genomes of these two viruses have the mature peptides annotated in GenBank. Therefore, the conserved mature peptide cleavage sites (9) were used to evaluate the accuracy of the VIGOR predictions. All mature peptide

sequences predicted by VIGOR were found to be accurate during inspection by human annotators (data not shown).

Parainfluenza viruses and Sendai virus

Parainfluenza viruses and Sendai virus are also members of the paramyxovirus family. They are negative-strand RNA viruses with genome sizes of 15-17 kb. hPIV-2 and hPIV-4 belong to the rubulavirus genus, whereas hPIV-1 and hPIV-3, bPIV-3 and Sendai virus constitute the respirovirus genus (21). Like other members of the paramyxovirus family, the P gene is also polycistronic and encodes two to six different proteins. Sendai virus is the best studied virus in the paramyxovirus family. The P gene gives rise to a maximum of six different proteins by means of RNA editing, overlapping frames and ribosomal shunting (5,22,23,12). The V protein is produced from the edited mRNA. The faithful transcript encodes the P protein. Four additional proteins (C', C, Y1 and Y2) are also encoded in the polycistronic P gene by a process known as ribosomal shunting (12,23). Here, different start codons including a non-ATG codon (ACG or GCG for C' protein in the UTR region of the P gene) are used by ribosome to initiate translations. In the hPIV-1 genome, the P gene doesn't encode the edited mRNA that will be translated into V or D protein as is the case in other related viruses (24). However, the small proteins (C', C, Y1 and Y2) encoded by the alternative start codons are detected in hPIV-1 genomes. The P gene of hPIV-3 encodes at least four proteins (P, D, C and V) (25). P and C proteins are encoded by the non-edited mRNA with different frames. The D protein will be translated from this edited mRNA with the insertion of two non-templated Gs at the RNA editing site (25). The edited mRNA of bPIV-3 encodes the D protein with two non-templated G residues inserted at the RNA editing site. A C protein is also encoded in the P gene with a different reading frame (26). hPIV-2 and hPIV-4 are members of the rubulavirus genus. Like other members in the same genus, the P protein is encoded by the edited mRNA with the addition of two G residues at the editing site, while the faithfully transcribed mRNA encodes the V protein (8,27). The genome size of hPIV-4 is ~17.7 kb, 2 kb longer than other parainfluenza virus genomes.

A total of 36 parainfluenza virus and Sendai virus genomes were run through VIGOR. Of the 311 genes predicted by VIGOR, 62 were not annotated in GenBank. The rest of VIGOR predictions completely agreed with GenBank annotations. The majority of VIGOR predictions that were not annotated in GenBank records are the C, Y1 and Y2 genes of Sendai virus, which overlap with the P gene but with different start codons and different reading frames.

Rubella virus

Rubella virus is the sole member of genus *Rubivirus* in family *Togaviridae*. It is a single-stranded, positive sense RNA virus. The genome size is ~9.8 kb with two genes encoded. The infections of rubella virus in pregnant women often result in severe defects in the embryos known as congenial rubella syndromes (28,29). The first gene encodes a large non-structural protein, which is cleaved into two mature peptides (P150 and P90) required in RNA replication. The second gene encodes a structural protein, which is cleaved into three mature peptides, the capsid protein (C) and two envelope glycoproteins (E1 and E2) (30). In order to identify the exact cleavage sites for the mature peptides, a mature peptide database of rubella virus was created. The polyproteins were BLASTed against the mature peptide database to locate the mature peptide regions. Conserved signature residues and alignment are used to find the exact cleavage sites. The conserved cleavage residues in the non-structural protein are Ser-Arg-Gly↓Gly-Gly (31).

A total of 46 complete genomes of rubella virus retrieved from GenBank were tested by VIGOR. All the 92 genes annotated in GenBank were detected by VIGOR, but the start codon of one gene was different (Table 1). Mature peptide sequences predicted by VIGOR were also validated by the GenBank annotations and by the analysis of the conserved cleavage sites (data not shown).

Respiratory syncytial virus

RSV still remains the most common disease agent of severe respiratory illness in child populations all over the world. Although a few RSV vaccines are available, RSV is still responsible for >90,000 hospitalizations in new borns and children each year in the USA (32). RSV are also members of the paramyxovirus family. RSV genome is ~15 kb and encodes 11 to 12 genes. Two regions in the RSV genome encode overlapping open reading frames. The first one is the matrix gene. A short, second open reading frame was identified in this region. The other overlapping gene region encodes transcription elongation factor M2-1 and M2-2 (33).

A total of 568 genes were predicted by VIGOR from 50 RSV genomes, and 522 genes predicted by VIGOR were the same as the annotations in GenBank. The start codons of 13 genes defined by VIGOR were different from GenBank annotations. All of these are M2-2 gene. Similar to the M2-1 protein in MPV, two versions of M2-2 protein exist in GenBank protein database. The longer version has two additional amino acids at the N-terminus. VIGOR picked the longer version as the reference sequence to define the start codon. VIGOR also predicted 33 genes that were not annotated in GenBank. All of these 33 new genes were the M2-2 and M1-2 genes.

Each of the different virus-specific VIGOR programs has also been tested with the viral genomic sequences generated by GSCID at JCVI. All the predictions have been examined by human annotators by comparing the predicted proteins with homologous proteins in GenBank. The evaluation data presented here indicated that VIGOR is not only able to accurately define the start and stop codons of genes in the designated viral genomes, but it can also precisely identify the complex gene features like RNA editing, stop codon read-through, alternative splicing and alternative translation initiation. Mature peptide prediction function has also been built into VIGOR for some viruses.

VIGOR was initially developed to annotate genomes of five different viruses (3). Here, we demonstrate that VIGOR has been extended to accurately predict protein coding genes and identify the complex gene features for additional 12 different viruses. If reference protein sequences of a new virus are available and the genome structure and complex gene features of the new virus are well understood, virus species-specific modules can be developed to identify the complex gene features (if they exist) in the virus genomes. With the drop of sequencing cost and the increase of sequencing productivity, tremendous number of different virus species from both clinical and environmental samples will be sequenced. The accuracy and efficiency of gene prediction will be critical to interpret newly sequenced viral genomes and to understand these viruses. VIGOR can be extended to annotate the genomes of new viruses by incorporating new reference sequences into the protein database and developing the virus species-specific modules.

ACKNOWLEDGEMENTS

The authors thank all the members of Viral Genome Informatics Affinity Group and Viral Genome Group for testing VIGOR and comments. Seth Schobel helped test the TBL file generation. The authors also acknowledge Tom Emmel and John Bury in the JCVI IT department for their help; the authors appreciate the valuable comments and suggestions provided by the two anonymous reviewers. VIGOR source codes will be available upon request.

FUNDING

NIAID, NIH, Department of Health and Human Services, federal funds [HHSN272200900007C, whole/in part]. Funding for open access charge: J. Craig Venter Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Guo,F.B. and Zhang,C.T. (2006) ZCURVE_V: a new self-training system for recognizing protein-coding genes in viral and phage genomes. *BMC Bioinformatics*, **7**, 9.
2. Bao,Y., Bolotov,P., Dernovoy,D., Kiryutin,B. and Tatusova,T. (2007) FLAN: a web server for influenza virus genome annotation. *Nucleic Acids Res.*, **35**, W280–W284.
3. Wang,S., Sundaram,J.P. and Spiro,D. (2010) VIGOR, an annotation program for small viral genomes. *BMC Bioinformatics*, **11**, 451.
4. Schneider,H., Kaelin,K. and Billeter,M.A. (1997) Recombinant measles viruses defective for RNA editing and V protein synthesis are viable in cultured cells. *Virology*, **227**, 314–322.
5. Vidal,S., Curran,J. and Kolakofsky,D. (1990) Editing of the Sendai virus P/C mRNA by G insertion occurs during mRNA synthesis via a virus-encoded activity. *J. Virol.*, **64**, 239–246.
6. Elliott,G.D., Yeo,R.P., Afzal,M.A., Simpson,E.J., Curran,J.A. and Rima,B.K. (1990) Strain-variable editing during transcription of the P gene of mumps virus may lead to the generation of non-structural proteins NS1 (V) and NS2. *J. Gen. Virol.*, **71 (Pt 7)**, 1555–1560.
7. Skiadopoulos,M.H., Vogel,L., Riggs,J.M., Surman,S.R., Collins,P.L. and Murphy,B.R. (2003) The genome length of human parainfluenza virus type 2 follows the rule of six, and recombinant viruses recovered from non-polyhexameric-length antigenomic cDNAs contain a biased distribution of correcting mutations. *J. Virol.*, **77**, 270–279.
8. Komada,H., Kawano,M., Uefuji,A., Ito,M., Tsurudome,M., Hatakeyama,E., Nakanishi,M., Sakue,S., Joh,C., Suzumura,E. *et al.* (2011) Completion of the full-length genome sequence of human parainfluenza virus types 4A and 4B: sequence analysis of the large protein genes and gene start, intergenic and end sequences. *Arch. Virol.*, **156**, 161–166.
9. Strauss,J.H. and Strauss,E.G. (1994) The alphaviruses: gene expression, replication, and evolution. *Microbiol. Rev.*, **58**, 491–562.
10. Latorre,P., Kolakofsky,D. and Curran,J. (1998) Sendai virus Y proteins are initiated by a ribosomal shunt. *Mol. Cell. Biol.*, **18**, 5021–5031.
11. Garcin,D., Curran,J., Itoh,M. and Kolakofsky,D. (2001) Longer and shorter forms of Sendai virus C proteins play different roles in modulating the cellular antiviral response. *J. Virol.*, **75**, 6800–6807.
12. Curran,J. and Kolakofsky,D. (1988) Ribosomal initiation from an ACG codon in the Sendai virus P/C mRNA. *EMBO J.*, **7**, 245–251.
13. Yun,S.I., Kim,S.Y., Rice,C.M. and Lee,Y.M. (2003) Development and application of a reverse genetics system for Japanese encephalitis virus. *J. Virol.*, **77**, 6450–6465.
14. von Lindern,J.J., Aroner,S., Barrett,N.D., Wicker,J.A., Davis,C.T. and Barrett,A.D. (2006) Genome analysis and phylogenetic relationships between east, central and west African isolates of Yellow fever virus. *J. Gen. Virol.*, **87**, 895–907.
15. Mead,P.S., Slutsker,L., Dietz,V., McCaig,L.F., Bresee,J.S., Shapiro,C., Griffin,P.M. and Tauxe,R.V. (1999) Food-related illness and death in the United States. *Emerg. Infect. Dis.*, **5**, 607–625.
16. Hardy,M.E. (2005) Norovirus protein structure and function. *FEMS Microbiol. Lett.*, **253**, 1–8.
17. Biacchesi,S., Skiadopoulos,M.H., Boivin,G., Hanson,C.T., Murphy,B.R., Collins,P.L. and Buchholz,U.J. (2003) Genetic diversity between human metapneumovirus subgroups. *Virology*, **315**, 1–9.
18. Lwamba,H.C., Alvarez,R., Wise,M.G., Yu,Q., Halvorson,D., Njenga,M.K. and Seal,B.S. (2005) Comparison of the full-length genome sequence of avian metapneumovirus subtype C with other paramyxoviruses. *Virus Res.*, **107**, 83–92.
19. Beier,H. and Grimm,M. (2001) Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. *Nucleic Acids Res.*, **29**, 4767–4782.
20. Li,G. and Rice,C.M. (1993) The signal for translational readthrough of a UGA codon in Sindbis virus RNA involves a single cytidine residue immediately downstream of the termination codon. *J. Virol.*, **67**, 5062–5067.
21. Henrickson,K.J. (2003) Parainfluenza viruses. *Clin. Microbiol. Rev.*, **16**, 242–264.
22. Kato,A., Kiyotani,K., Sakai,Y., Yoshida,T. and Nagai,Y. (1997) The paramyxovirus, Sendai virus, V protein encodes a luxury function required for viral pathogenesis. *EMBO J.*, **16**, 578–587.
23. de Breyne,S., Simonet,V., Pelet,T. and Curran,J. (2003) Identification of a cis-acting element required for shunt-mediated translational initiation of the Sendai virus Y proteins. *Nucleic Acids Res.*, **31**, 608–618.
24. Matsuoka,Y., Curran,J., Pelet,T., Kolakofsky,D., Ray,R. and Compans,R.W. (1991) The P gene of human parainfluenza virus type 1 encodes P and C proteins but not a cysteine-rich V protein. *J. Virol.*, **65**, 3406–3410.
25. Durbin,A.P., McAuliffe,J.M., Collins,P.L. and Murphy,B.R. (1999) Mutations in the C, D, and V open reading frames of human parainfluenza virus type 3 attenuate replication in rodents and primates. *Virology*, **261**, 319–330.

26. Jacques,J.P., Hausmann,S. and Kolakofsky,D. (1994) Paramyxovirus mRNA editing leads to G deletions as well as insertions. *EMBO J.*, **13**, 5496–5503.
27. Nishio,M., Tsurudome,M., Ito,M., Garcin,D., Kolakofsky,D. and Ito,Y. (2005) Identification of paramyxovirus V protein residues essential for STAT protein degradation and promotion of virus replication. *J. Virol.*, **79**, 8591–8601.
28. Tzeng,W.P. and Frey,T.K. (2005) Rubella virus capsid protein modulation of viral genomic and subgenomic RNA synthesis. *Virology*, **337**, 327–334.
29. Zhou,Y., Ushijima,H. and Frey,T.K. (2007) Genomic analysis of diverse rubella virus genotypes. *J. Gen. Virol.*, **88**, 932–941.
30. Yao,J., Yang,D., Chong,P., Hwang,D., Liang,Y. and Gillam,S. (1998) Proteolytic processing of rubella virus nonstructural proteins. *Virology*, **246**, 74–82.
31. Chen,J.P., Strauss,J.H., Strauss,E.G. and Frey,T.K. (1996) Characterization of the rubella virus nonstructural protease domain and its cleavage site. *J. Virol.*, **70**, 4707–4713.
32. Anderson,L.J., Parker,R.A. and Strikas,R.L. (1990) Association between respiratory syncytial virus outbreaks and lower respiratory tract deaths of infants and young children. *J. Infect. Dis.*, **161**, 640–646.
33. Karron,R.A., Buonagurio,D.A., Georgiu,A.F., Whitehead,S.S., Adamus,J.E., Clements-Mann,M.L., Harris,D.O., Randolph,V.B., Udem,S.A., Murphy,B.R. *et al.* (1997) Respiratory syncytial virus (RSV) SH and G proteins are not essential for viral replication in vitro: clinical evaluation and molecular characterization of a cold-passaged, attenuated RSV subgroup B mutant. *Proc. Nat. Acad. Sci. U.S.A.*, **94**, 13961–13966.